Evolution of antiparallel two-domain membrane proteins

Lolkema, Juke S.; Dobrowolski, Adam; Slotboom, Dirk-Jan

Published in:
Journal of Molecular Biology

DOI:
10.1016/j.jmb.2008.03.005

Publication date:
2008

doi:10.1016/j.jmb.2008.03.005 J. Mol. Biol. (2008) 378, 596–606

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Evolution of Antiparallel Two-domain Membrane Proteins: Tracing Multiple Gene Duplication Events in the DUF606 Family

Juke S. Lolkema[1]*, Adam Dobrowolski[1] and Dirk-Jan Slotboom[2]

[1]*Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands*

[2]*Membrane Enzymology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands*

Received 18 January 2008; received in revised form 29 February 2008; accepted 4 March 2008
Available online 12 March 2008

X-ray crystallography has revealed that many integral membrane proteins consist of two domains with a similar fold but opposite (antiparallel) orientation in the membrane. The proteins are believed to have evolved by gene duplication and gene fusion events from a dual topology ancestral membrane protein, that adapted both orientations in the membrane and formed antiparallel homodimers. Here, we present a detailed analysis of the DUF606 family of bacterial membrane proteins that contains the entire collection of intermediate states of such an evolutionary pathway: single genes that would code for dual topology homodimeric proteins, paired genes coding for homologous proteins with a fixed but opposite orientation in the membrane that would form heterodimers, and fused genes that encode antiparallel two-domain fusion proteins. Two types of paired genes can be discriminated corresponding to the order in which the genes coding for the two oppositely oriented proteins occur in the operon. On the protein level, the heterodimers resulting from the two types of gene pairs are indistinguishable. In contrast, two types of fused genes corresponding to the two possible orders in which the oppositely oriented domains are present in the encoded proteins, do result in discernible types of proteins. The large number of genetic and protein states in the DUF606 family allowed for a detailed phylogenic analysis that revealed a total of nine independent duplication events in the DUF606 family, five of which resulted in paired genes, and four resulted in fused genes. Noticeably, there was no evidence for a sequential mechanism in which fusions evolve from a pair of genes. Rather, an evolutionary mechanism is proposed by which antiparallel two-domain proteins are the direct result of a gene duplication event. Combining the phylogeny of proteins and hosting microorganisms allowed for a reconstruction of the evolutionary pathway.

Keywords: dual topology; DUF606; evolutionary state; gene duplication; gene fusion

Edited by J. Bowie

Introduction

Crystal structures have revealed that many membrane proteins consist of two domains that share a similar fold, most likely the result of an ancient gene duplication. The two domains in membrane proteins have the same (parallel) or opposite (antiparallel) orientation in the membrane corresponding to an even or odd number of transmembrane segments per domain. The antiparallel domain organization is observed more frequently in the 3D structures of membrane proteins than the parallel domain organization. The aquaporins,[1,2] the ammonia channel AmtB,[3] the bacterial preprotein translocase subunit SecY,[4] the Na^+-leucine transporter LeuT,[5] the CLC chloride transporters/channels,[6] the Na^+-H^+ antiporter,[7] and the ABC transporter subunit BtuC,[8] all contain two easily recognizable domains or, at least, two structural elements that have the same fold and that are oriented oppositely in the membrane. In addition, biochemical evidence has been presented in-

*Corresponding author. E-mail address: j.s.lolkema@rug.nl.

Abbreviation used: TMS, transmembrane segment.

dicating this structural organization to be more widespread.[9–11] Although the domain structure is clearly recognizable in the high-resolution crystal structures, the homology was in most cases not obvious from the amino acid sequences of the two domains because they have diverged too far.

To explain how the membrane proteins with two domains of antiparallel orientation may have arisen during evolution, the existence of evolutionary ancestral proteins with dual topology has been postulated, membrane proteins that insert into the membrane in both orientations.[12–14] The dual topology ancestor protein would associate into a homodimer with the two identical subunits having opposite orientation in the membrane. A gene duplication followed by divergence resulted in a heterodimeric protein with subunits of fixed but opposite membrane orientation. Eventually, a fusion of the two genes yielded the two-domain proteins as we see them today.

Experimental support for such an evolutionary pathway comes from studies of members of the small multidrug-resistant (SMR) transporter family. The EmrE protein in this family is coded by a single gene in the genome of *Escherichia coli* and assembles into a homodimer in the membrane. Though heavily debated,[15] evidence has been presented indicating that the two subunits have opposite orientation in the membrane,[14,16] thereby supporting dual topology. In the same family, *ebrA* and *ebrB* are a pair of genes forming an operon on the *Bacillus subtilis* chromosome. The gene products assemble into a heterodimer, of which the two subunits were shown to have anti-parallel orientation.[17] Bioinformatics studies have provided further support for the proposed evolutionary pathway by identifying a number of protein families with apparent successive intermediate states in different organisms.[14] These studies were based on the analysis of the positive charge bias in the loops on either side of the membrane (positive-inside rule).[18,19] One family, termed DUF606, appeared to be especially rich in evolutionary states. The genes were found as single entities (singletons), in pairs or as two-gene fusions. The proteins coded by the single genes showed very little or no positive charge bias, suggestive of dual topology. The proteins coded by the paired genes showed a significant charge bias, and each pair coded consistently for oppositely oriented proteins. The proteins coded by the fused genes contained two homologous domains with opposite orientation.[14] Here, we give a detailed analysis of the proteins of the DUF606 family with the aim to reconstruct the evolutionary pathway(s) by which the diversity has arisen.

Results

The DUF606 family: singles, pairs and fusions

A total of 369 genomes of microbial species available on June 1, 2007 contained 148 DUF606 family members (Supplementary Data Table A). No member could be detected in the eukaryotic domain. The DUF606 family contains almost exclusively bacterial membrane proteins; only one member was found in the genome of an archeon, the euryarchaeota *Methanococcus maripaludis*. The gene was identified in the S2, C5 and C7 strains, which makes an erroneous annotation unlikely. The 148 members were species-specific rather than strain-specific; only eight were found in one strain but not in other strains of a species (Supplementary Data Table A). The DUF606 proteins are quite common in bacteria. They are found in roughly half of the Firmicutes and are somewhat less abundant in Proteobacteria (~40%) and Actinobacteria (~30%) (Supplementary Data Table B). Inspection of the neighborhood of the 148 genes on the chromosomes showed that 73 were present as single genes coding for a protein of around 150 amino acid residues, 52 were present in pairs of adjacent genes each coding for similarly sized proteins, and 23 genes contained an internal duplication (fusions) coding for proteins of twice the length (see below). Singles and pairs were found mostly in Firmicutes and Proteobacteria, while fusions were particularly successful in Firmicutes and Actinobacteria.

Membrane topology

A consensus membrane topology model of the DUF606 proteins was constructed by averaging the TMHMM predictions (see Computational Methods) over all members using a multiple sequence alignment to align the different positions in the models (Fig. 1). The consensus model suggests the presence of five transmembrane segments (TMSs) in the proteins coded by the genes in the singles and the pairs groups, and ten TMSs in the fusions. The latter appear to be organized in two groups of five TMSs each. The internal gene duplication in the fusions group was confirmed by splitting the encoded proteins into two halves around position 225 in the multiple sequence alignment (see Fig. 1c) and aligning the two halves. Sequence identity between the N- and C-terminal halves ranged between 20% and 42%, showing clearly that the two halves represent two homologous domains with five predicted TMSs each. For comparison, sequence identity between the two proteins coded by the paired genes ranged from 20–53%. Since the number of TMSs in each domain of the fusions is odd, the domains are predicted to have the opposite orientation in the membrane.[14] The position of the predicted TMSs in the proteins coded by the singles, pairs and fusions are indicated in the top of the graphs in Fig. 1. TMHMM does not give a consistent prediction about the orientation of the proteins in the membrane for any of the three groups, and therefore the orientation is not indicated. In the following discussion, all the hydrophilic regions situated at the same side of the membrane as the N terminus are referred to as the Odd-loops, and the hydrophilic regions at the other side of the membrane as the Even-loops.

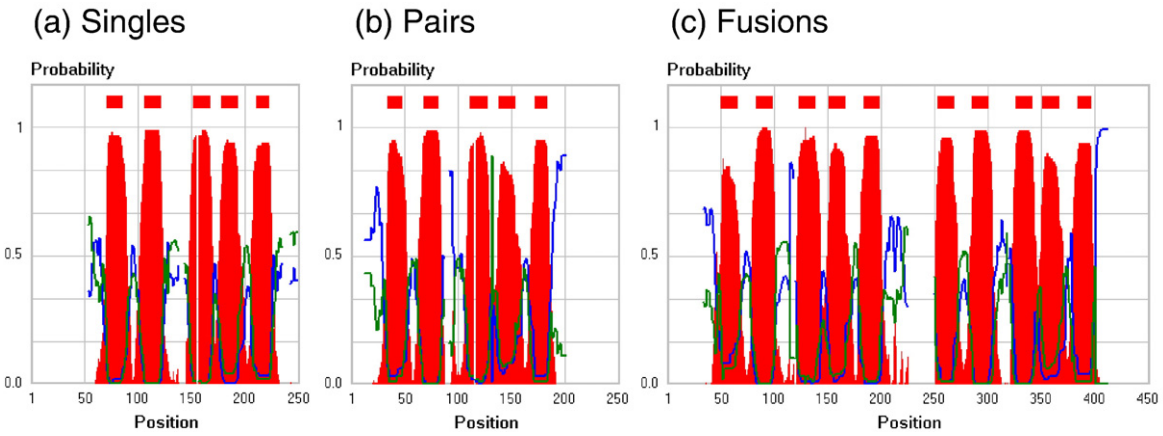

Fig. 1. Consensus membrane topology models of singles (a), pairs (b), and fusions (c) in the DUF606 family. Averaged predictions are shown for transmembrane regions (red), cytoplasmic loops (blue), and outside loops (green). Red blocks at the top of the graphs indicate the position of the predicted transmembrane segments in the model.

Charge distribution in the loops

The distribution of positively charged amino acid residues in the loop regions of the topology models was analyzed by plotting the number of positive charges in the Odd-loops against the number in the Even-loops for each protein (Fig. 2). The data points for the proteins coded by the single genes cluster around the diagonal, showing that there is little positive charge bias in either the Odd- or Even-loops. Therefore, according to the positive inside rule, the proteins do not seem to be directed into one particular orientation in the membrane, which is consistent with their proposed dual topology behavior. The proteins coded by the genes present in pairs clearly separate into two groups, one with the data points above the diagonal, the other below. The number of proteins in the two groups is the same, and for each pair one protein is represented in the group above the diagonal and the other in the group below. The group below the diagonal has a surplus of positive charges in the Odd-loops, indicating that the N terminus of the proteins is located in the cytoplasm. Conversely, the positive charge bias is towards the Even-loops in the group above the diagonal, indicating that the orientation of the proteins in the membrane is opposite and the N terminus is on the outside. It follows that in each pair one protein is predicted to have the N terminus in the cytoplasm (the up orientation) and the other is predicted to have the N terminus externally (the down orientation). The proteins coded by the fused genes also fall into two groups (Fig. 2c). Eight cluster below the diagonal, indicating that they insert into the membrane with their N termini in the cytoplasm, the remaining 15 proteins cluster above the diagonal, indicating that they have the opposite orientation.

Evolutionary states

The genes that come in pairs are coded by the same DNA strand and predicted to be organized in an

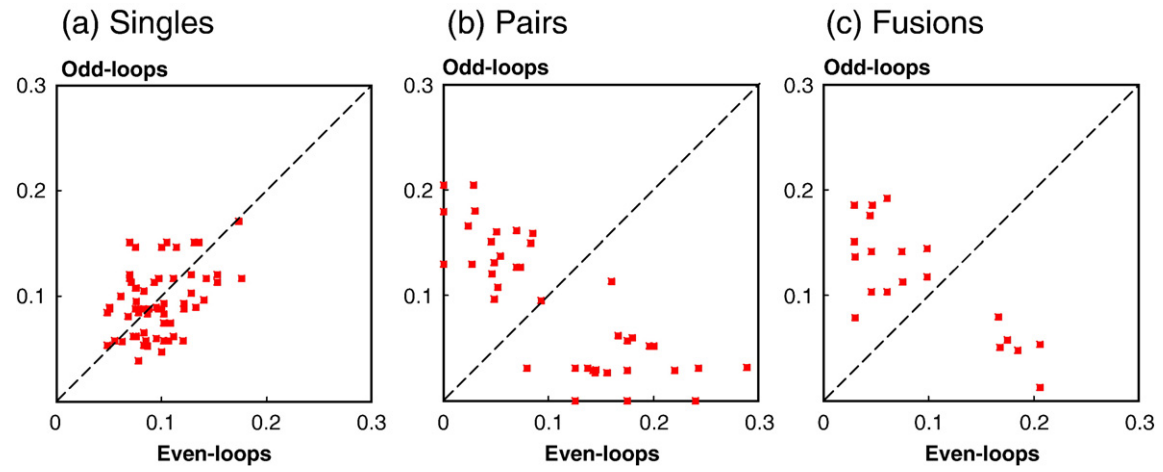

Fig. 2. Distribution of positive charges (K+R) over odd and Even-loops in the consensus topology model of singles (a), pairs (b), and fusions (c) in the DUF606 family. In the graph, the positive charge density (defined in Computational Methods) in the Odd-loops was plotted against the positive charge density in the Even-loops.

operon structure. We refer to the genes as either the first (at the 5′ position in the operon) or the second (at the 3′ position in the operon). There are two types of pairs; 26 with the first gene coding for the subunit in the up orientation and four in which the order is reversed (see also Table 1). This situation resembles that of the fused genes where the two domains have the up/down or the down/up configuration (Fig. 2c). In contrast to the fusions, the different order of the genes in the pairs has no consequences at the protein level but is relevant from an evolutionary point of view. Figure 3 shows the relation between the different evolutionary states at the genetic and protein levels. There are five genetic states: the single gene, two types of gene pairs, and two types of gene fusions. The five genetic states result in four protein states. Singles result in homodimers, both types of pairs in heterodimers that are indistinguishable and the two types of fusions in two types of two-domain proteins with the two possible orders of the domains.

Gene duplication events

A phylogenetic tree of the members of the DUF606 family that were found in pairs revealed several well-separated clades, each consisting of proteins encoded by genes found at the same position in the operon, either the first (blue) or the second (green) position (Fig. 4). Five clades with proteins encoded by the first gene in the operon may be discriminated (P1–5, blue) and five with proteins encoded by the second gene (P1–5, green). Noticeably, if the proteins coded by the first gene in the operon cluster together in a monophyletic clade (e.g. P3, green), the corresponding partners from the second position also cluster together (e.g. P3, blue). The pairs of the down/up type are represented exclusively in subgroup P1, and all the pairs of the up/down type are in subgroups P2–5 (Table 1). Tracing back the two clades of each subgroup to the node where they merge reveals the origin of duplication of the pairs in the subgroup. The two clades of subgroup P1 (blue and green) together form a new monophyletic clade that is separated from all the other clades in the tree (bootstrap confidence of 100%). Because only subgroup P1 contains pairs of the down/up type, it is clear that down/up pairs are the result of a different duplication event than the event(s) that gave rise to the up/down pairs. By the same token, also subgroup P3 clearly has arisen from a separate duplication event, indicating that pairs of the up/down type result from more than one duplication event. A systematic pairwise phylogenetic analysis of all the subgroups is consistent with each of the subgroups P1–5 being the result of separate duplication events. An example in which subgroups P3 and P5 are compared is given in Fig. 5b. The nodes where the two clades of the subgroups merge in the tree (indicated by the arrows) are separated with a bootstrap confidence of 98%. The data for all combinations are presented in Supplementary Data Table C.

In the phylogenetic tree of the fusions, the eight up/down proteins are well separated from the 15 down/up proteins (bootstrap significance of 100%; Supplementary Data Fig. S1) suggesting separate

Table 1. Phylogenetic groups and orientation of pairs

Group	Lineage[a]	Orientation	Organism	Pair
P1	B-F-b	Down/up	*Listeria monocytogenes*	23651220lmon/23651220lmon
			Listeria innocua	LIN1174linn/LIN1173linn
			Listeria welshimeri	LWE1168lwel/LWE1167lwel
			Bacillus licheniformis	BLI01976blic/BLI01975blic
P2	B-F-b	Up/down	*Bacillus cereus*	BC2334bcer/BC2336bcer
P3	B-P-c	Up/down	*Enterobacter sp.*	T6381093ensp/T6381092ensp
			Acinetobacter baumannii	YP00108483abau/A1S1802abau
			Pseudomonas fluorescens	PFL2264pflu/PFL2265pflu
			Pseudomonas fluorescens	PFL3766pflu/PFL3765pflu
			Chromohalobacter salexigens	CSAL0823csal/CSAL0822csa
			Psychrobacter arcticus	PSYC0049parc/PSYC0048parc
			Psychrobacter cryohalolentis	CRYO1328pcry/CRYO1329pcry
P4	B-F-b	Up/down	*Bacillus thuringiensis*	97273229bthu/97273230bthu
			Bacillus anthracis	BAS3256bant/BAS3257bant
			Bacillus cereus	BCZK3166bcer/BCZK3167bcer
	B-P-a	Up/down	*Agrobacterium tumefaciens*	ATU2062atum/ATU2063atum
			Rhizobium leguminosarum	RL110473rleg/RL110474rleg
			Rhizobium etli	EPE00364retl/EPE00365retl
	B-P-b	Up/down	*Chromobacterium violaceum*	CV3483cvio/CV3484cvio
	B-P-c	Up/down	*Photorhabdus luminescens*	PLU2183plum/PLU2184plum
			Shewanella loihica	YP00109441sloi/YP00109441_1sloi
			Shewanella oneidensis	SO0370sone/SO0371sone
			Shewanella sp.	ANA33799shsp/ANA33798shsp
P5	B-Y	Up/down	*Nostoc sp.*	ALL7167nsp./ALL7166nsp.
			Synechococcus elongatus	79421161selo/79421162selo
	B-D		*Deinococcus radiodurans*	DR1112drad/DR1113drad

[a] First character: B, Bacteria. Second character: F, Firmicutes; P, Proteobacteria; Y, Cyanobacteria; D, Deinococcus; A, Actinobacteria. Third character: F-b, Bacillales; l, Lactobacillales; B-c, Clostridia; a, α-subdivision; P-b, β-subdivision; P-c, γ-subdivision; and d, δ-subdivision.

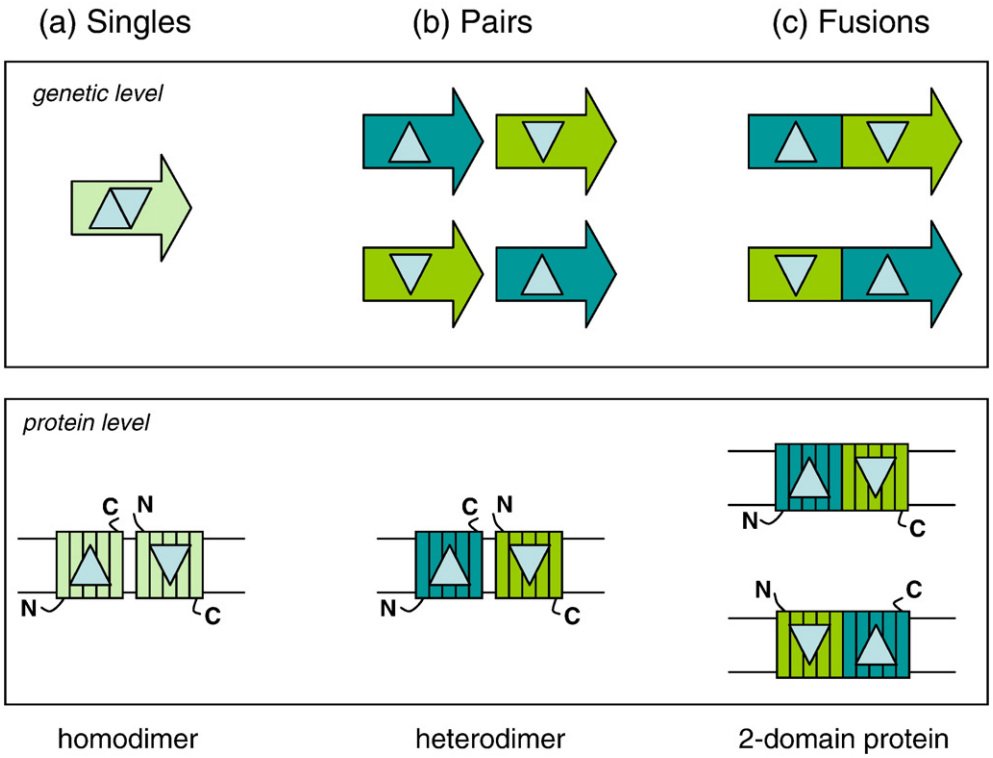

Fig. 3. Evolutionary states of DUF606 members on the genetic (top) and the protein (bottom) level. Top panel: Arrows indicate genes. Triangles and inverted triangles indicate the orientation of the encoded proteins in the membrane as up (blue) or down (green) or dual (pale green). Bottom panel: Boxes indicate the DUF606 proteins/domains with five TMSs each in either the up or down topology. See the text for an explanation.

duplication events. A phylogenetic tree of the separated domains of the fusion proteins was made to analyze the duplication events that led to the fusion proteins in detail (Fig. 6). Similar to the tree of the pairs, several well-separated clades may be discriminated. Each clade contains either exclusively N-terminal domains, or C-terminal domains. There are four clades with N-terminal domains (F1–4, blue) and four clades with C-terminal domains (F1–4, green). Without exception, whenever a group of N-terminal domains forms a separate clade, the corresponding C-terminal domains form a separate clade. The up/down fusions are found in subgroup F1, the down/up fusions in F2–4 (Table 2). Using the same strategy of comparing the different subgroups with each other as described above for the pairs, it follows that the four subgroups represent four different duplication events (Supplementary Data Table C). An example is shown in Fig. 5a.

It is concluded that at least five duplication events have taken place in the group of DUF606 members that come in pairs and four in the group that come as fusion proteins.

Single duplication *versus* sequential mechanism

One possible model for the evolution of the two-domain membrane proteins (fusions) assumes two sequential events. In the first event, an ancestral gene duplicated to yield a pair of genes and, subsequently, in the second event the paired genes fused to encode the fusion protein. This model implies that fusion proteins and paired proteins of the DUF606 family that we observe in present day organisms may have arisen from the same duplication event. In a second model, the two-domain protein is the result of just a single event: duplication of an ancestral gene resulted either in a pair of genes or, directly, in fused genes. This model implies that the fusion proteins and paired proteins in present-day organisms have arisen from different duplication events. The two models lead to distinguishable phylogenetic trees: In the two-event model, the N-terminal domains of fusion proteins and the proteins from the pairs that are coded by the genes in the first positions of the operons cluster into the same clade. Similarly, the C-terminal domains cluster with the proteins from the second positions. On the other hand, in the one-event model, the N and C-terminal domains of the fusion proteins would always cluster together rather than blend with proteins that come in pairs. Analysis of the phylogenetic relation of the five subgroups of the pairs and the four subgroups of the fusions did not show any mixing of the clades from the fusion and the paired proteins (Supplementary Data Table C). As an example, the tree comparing fusions F1 and pairs P3 indicates that the two represent separate duplication events, with bootstrap confidence of 100% (Fig. 5c). It is more likely, therefore, that the formation of pairs and fusions in the DUF606

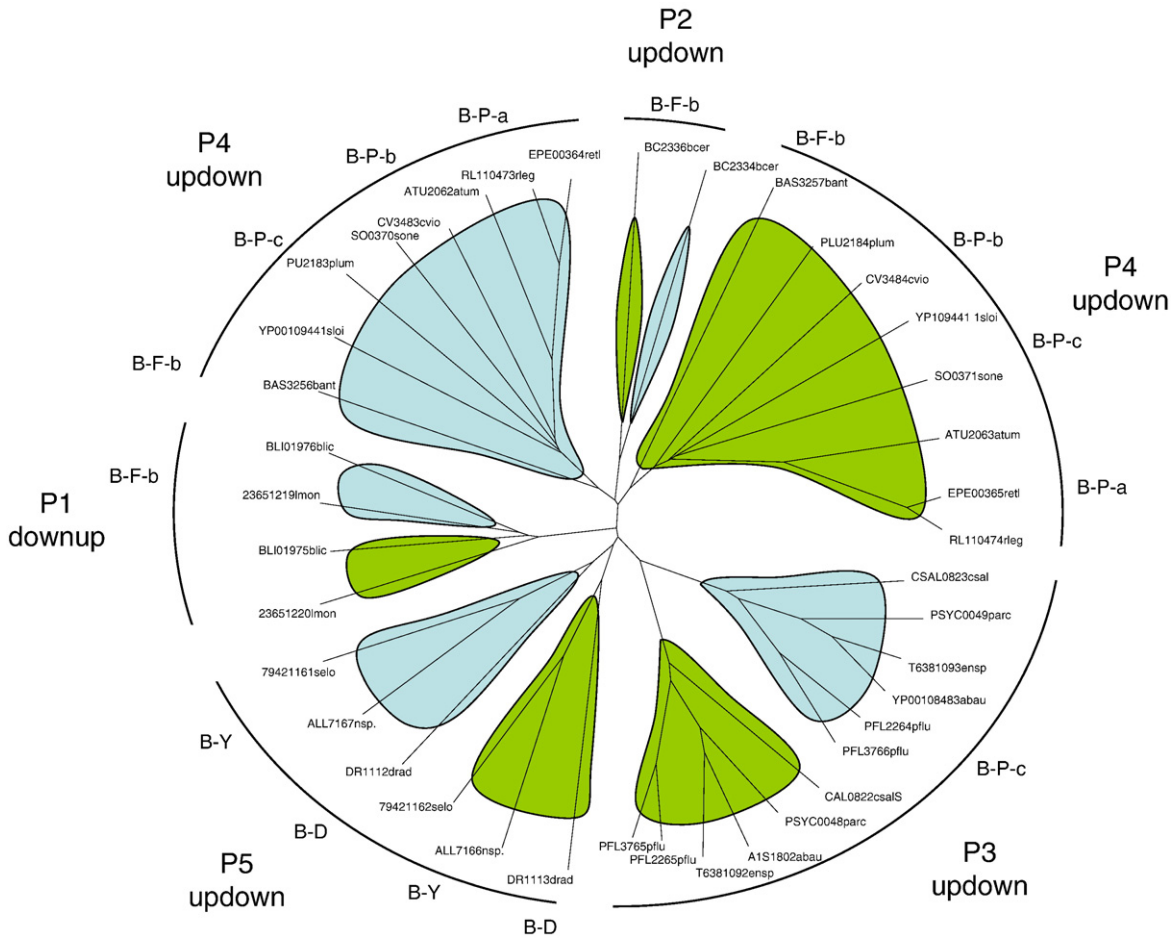

Fig. 4. Phylogenetic tree of the pairs. Subgroups P1 through P5 are indicated on the tree as well as the phylogenetic lineages of the hosting organisms within the groups (see Table 1). The position of the encoding genes in the pair on the chromosome are in blue for the first position, and in green for the second position. See the text for interpretation.

family of proteins are the result of single evolutionary events (see Discussion).

Discussion

Mechanism of evolution of two-domain membrane proteins

Our phylogenetic analysis of the DUF606 family of membrane proteins indicates that multiple duplication events have taken place in the family, at least five of which led to the formation of gene pairs coding for two proteins with opposite orientation in the membrane, and another four of which created fusion proteins. In the set of sequences available, no indication for sequential events, in which a duplication is followed by a fusion, was found. Mixed clades of paired proteins and the domains of fusion proteins that would have been evidence for sequential events were not observed in the trees. Obviously, the absence of mixed clades does not formally exclude that sequential events took place, but it would require additional assumptions to explain why they are silent in the analysis. For instance, it is possible that fusion of all pairs in the same clade took place in parallel, or that the duplication and fusion events both took place in an ancestral organism before it evolved further into different organisms. Both possibilities appear unlikely and the most parsimonious explanation of the data is that sequential events have not taken place. At any rate, there is no need to postulate a sequential mechanism, as will be discussed next.

A duplication of the ancestral gene coding for a dual topology protein is likely to be successful only if it results in a single transcript coding for both subunits. This warrants that the subunits are produced at the same time to form a complex and may facilitate production with the correct stoichiometry. A single transcript is obtained when the start codon of the second gene is close to the stop codon of the first gene. The former may be upstream of the latter, resulting in overlapping genes or immediately downstream. Analysis of the intergenic region of the DUF606 family members in the pairs groups showed that the distance between stop and start of the two genes ranged from −16 to +22 base-pairs, i.e. the genes are in very close vicinity (Supplementary Data Table D). A fusion is a special case when the start codon of the second gene is upstream of the stop

codon and the reading frames of the two genes are in frame. Statistically, the chances for a duplication to result in a fusion are considerably lower, but this may be compensated by a selective advantage in protein biogenesis, especially in translation because separate ribosomal binding sites are not needed and production of equal amounts of the two domains is guaranteed. The more or less similar frequencies of pairs and fusion observed in this study appear to support this analysis.

Gene duplications resulting in larger distances between the two genes are likely to evolve further independently towards different functions. However, very few paralogues are found in the DUF606 family. Moreover, they are found on separate clades in the family tree, indicating that they have a higher sequence identity with members in other organisms than to each other, suggesting the involvement of horizontal gene transfer. Apparently, gene duplications in the DUF606 family other than resulting in pairs or fusions is not an advantageous event for the organism.

Following a gene duplication resulting in either a pair or a fusion, sequence identity between the two encoded subunits/domains will be high at first but, as the defined orientation in the membrane develops, it will decrease gradually. All the pairs and fusion proteins in the DUF606 family that are found in contemporary sequenced organisms appear to have developed a fixed orientation (Fig. 2). Apparently, a situation in which a pair of dual topology proteins is present, or a fusion protein with dual topology is a selective disadvantage. The gene pairs are found in the up/down or down/up order, and the fusions are either up/down or down/up (see Fig. 3). There is no *a priori* reason why the first or second domain in a fusion, or why proteins encoded by the first or second gene in the operon, should have the up or down orientation in the membrane. Following the duplication, the evolution into the up/down and down/up states could go in either direction. However, intriguingly, duplication events that resulted in pairs evolved four times more frequently in up/down pairs than in down/up pairs. For fusions; this is the other way around; three of the duplications result in down/up fusions and only one is up/down. Possibly, there is a selective advantage for up/down pairs and for down/up fusions, but it is not clear what would be the selection criterion. It must be noted that the numbers are small and that what we see may be random fluctuation only.

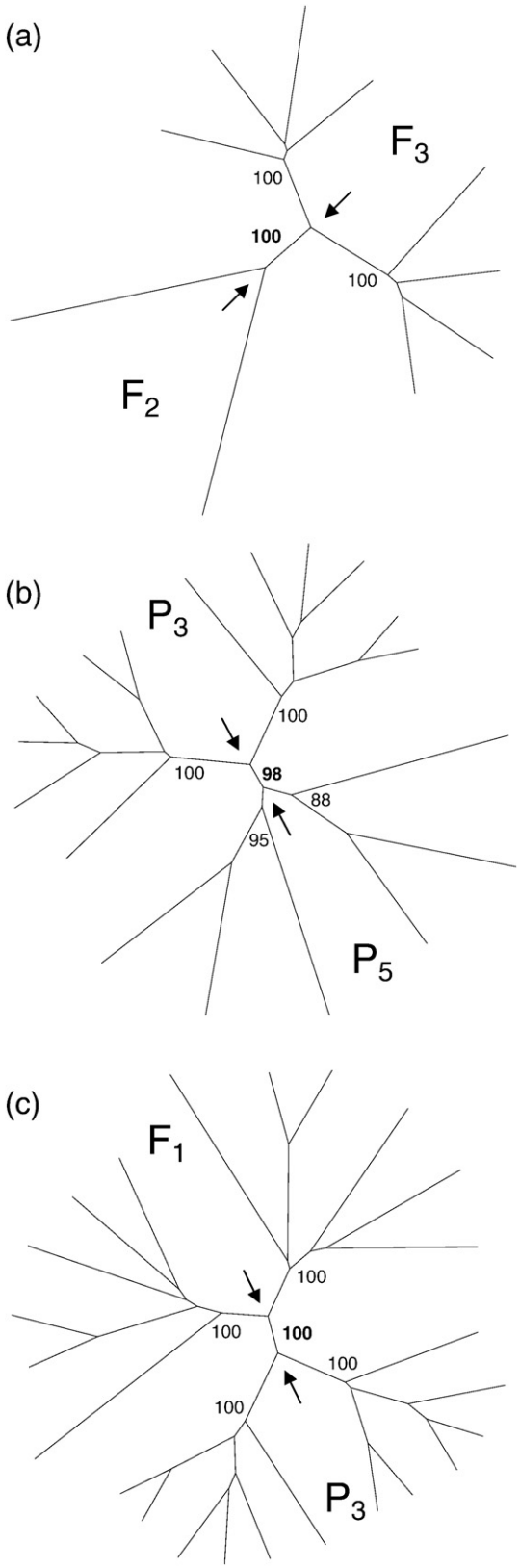

Fig. 5. Phylogenetic trees of domains of fusion proteins from subgroups F2 and F3 (a), of pairs from P3 and P5 (b), and of domains and pairs from F1 and P3 (c). The bootstrap values indicating the separation of the two subgroups are shown in bold; those indicating the clustering of the corresponding sequences within one subgroup of pairs or domains are shown in normal font. The trees demonstrate that the domains or pairs within one subgroup are more related to each other than to domains or pairs in the other subgroup, suggesting separate duplication events (indicated by the arrows).

Reconstruction of the evolutionary pathway

Pairs in subgroups P1, P2, and P3 and fusions in subgroups F2 and F3 are found in bacteria from well defined phylogenetic niches (see Tables 1 and 2),

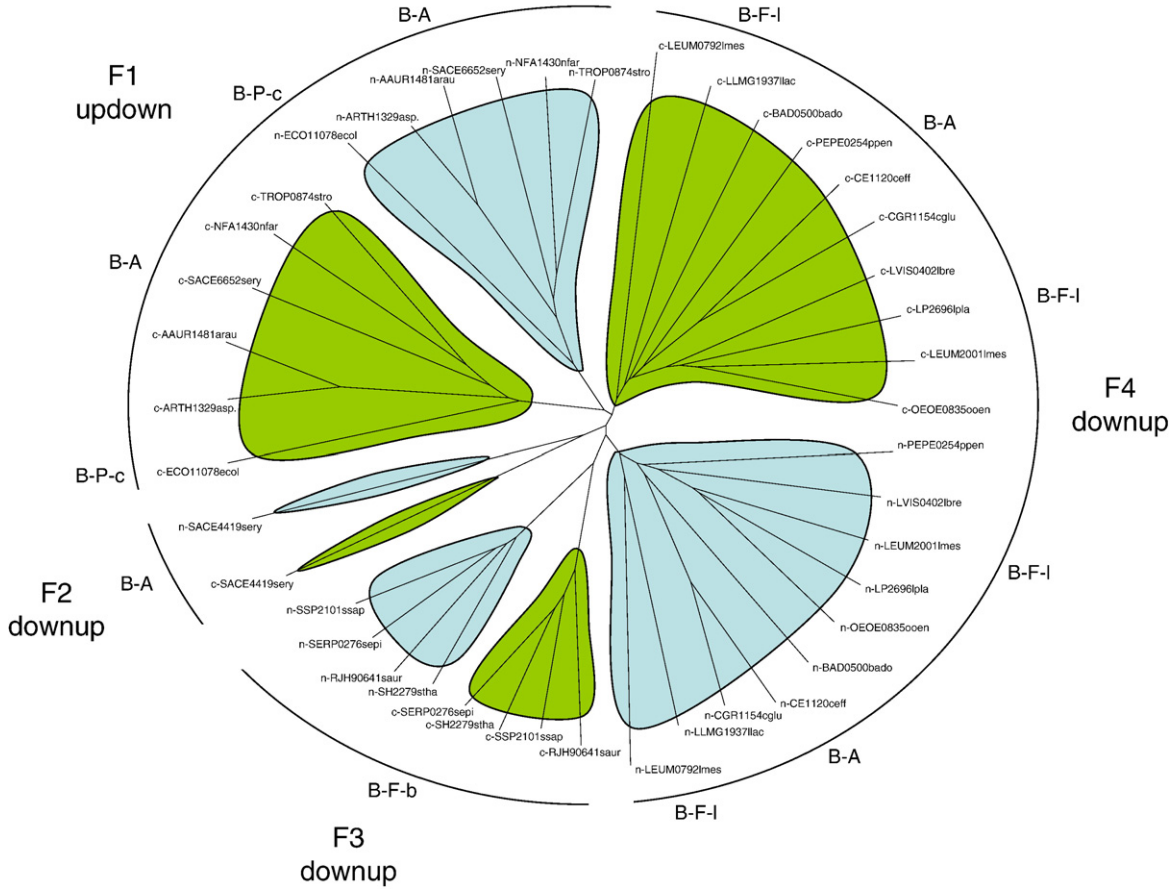

Fig. 6. Phylogenetic tree of the domains of the fusion proteins. The fusion proteins indicated in Table 2 were split into sequences coding for the two domains followed by multiple sequence alignment. Subgroups F1 through F5 are indicated on the tree as well as the phylogenetic lineages of the hosting organisms within the groups (see Table 2). The position of the domains in the fusion proteins are in blue for N-terminal domains and in green for C-terminal domains.

indicating that the pairs and fusions evolved in a single lineage. In contrast, pairs of subgroups P4 and P5, and fusions of F1 and F4 are phylogenetically more diverse containing proteins from different phyla. Figure 7 presents a reconstruction of the duplication and HGT events in the evolution of the DUF606 family projected onto the evolution of the organisms. The model is restricted to Proteobacteria, Firmicutes and Actinobacteria since these represent the majority of the dataset.

Proteobacterial DUF606 members are found in subgroups P3, P4 and F1, as well as in the group of singles. The latter are the most abundant in Proteobacteria and propagated in the α, β, γ and δ-subdivision (Supplementary Data Table A). Therefore, dual topology proteins appear to be the most prominent type of DUF606 proteins among Proteobacteria. Pairs of subgroup P4 are found in the α-, β-, and γ-subdivisions, indicating that the corresponding duplication event took place before the Proteobacteria evolved into subdivisions. For comparison, pairs of subgroup P3 are found in the γ-subdivision only indicating the event took place in a primordial γ-Proteobacterium after the Proteobacteria split up in subdivisions. In line with this, the proteins in pairs P4 and P3 share 20–26% and 34–38% sequence identity, respectively, indicating that the P4 pairs represent a more ancient duplication event. Pairs P4 are also observed in three organisms that do not belong to the Proteobacteria, the *Bacillus* species, *cereus, anthracis and thuringiensis* in phylum Firmicutes. The restricted distribution over three closely related bacteria together with the similar evolutionary distance of the two sequences of the pairs (21%) suggests that they are the result of a cross-phylum HGT event carrying a pair from a Proteobacterium to a Bacillus species. Fusion proteins are rare in Proteobacteria (subgroup F1) and most likely the result of HGT (see below).

Firmicutes contain a diverse set of DUF606 members with proteins from the group of singles and the subgroups F3, F4, P1, P2 and P4. In contrast to the Proteobacteria, singles in Firmicutes propagated in a single class only, in *Clostridia*. Fusions F4 are specific for class Lactobacillales and found in many different genera suggesting that the duplication event took place at the root of this class. In contrast, fusions F3 are found in a very limited niche, in four different *Staphylococci* and, therefore, the result of a more recent duplication event that took place at the level of the genera. In agreement, sequence identity shared by the two domains of the fusions in the *Staphylo-*

Table 2. Phylogenetic groups and domain orientation of fusions

Group	Lineage[a]	Orientation	Organism	Protein
F1	B-P-c	Up/down	*Escherichia coli*	ECO11078eco
			Salmonella typhimurium	STM3549styp
			Salmonella enterica	SC3479saen
	B-A	Up/down	*Salinispora tropica*	TROP0874stro
			Saccharopolyspora erythraea	SACE6652sery
			Arthrobacter aurescens	AAUR1481arau
			Nocardia farcinica	NFA1430nfar
			Arthrobacter sp.	ARTH1329asp.
F2	B-A	Down/up	*Saccharopolyspora erythraea*	SACE4419sery
F3	B-F-b	Down/up	*Staphylococcus haemolyticus*	SH2279stha
			Staphylococcus saprophyticus	SSP2101ssap
			Staphylococcus epidermidis	SERP0276sepi
			Staphylococcus aureus	RJH90641saur
F4	B-F-l	Down/up	*Pediococcus pentosaceus*	PEPE0254ppen
			Oenococcus oeni	OEOE0835ooen
			Leuconostoc mesenteroides	LEUM2001lmes
			Lactobacillus plantarum	LP2696lpla
			Lactobacillus brevis	LVIS0402lbre
			Lactococcus lactis	LLMG1937llac
			Leuconostoc mesenteroides	LEUM0792lmes
	B-A	Down/up	*Corynebacterium glutamicum*	CGR1154cglu
			Corynebacterium efficiens	CE1120ceff
			Bifidobacterium adolescentis	BAD0500bado

[a] See the legend to Table 1 for abbreviations.

coccus species is high, ranging between 37% and 42%, as compared to 25–34% in the Lactobacillales proteins in F4. Like fusions F3, pairs P1 are the result of a recent duplication event. The pairs are found in *Listeria* species and the sequence identity between the proteins is high (48–53%). Pair P2 is interesting because it is found only in one particular strain of *Bacillus cereus* and not in three others, and it is on the same clade in the tree of DUF606 members as a group of four singles found in *Clostridia* species (not shown), suggesting a common ancestor. Possibly, a gene coding for a single in class Clostridia was transferred laterally to the Bacillus species in which it has duplicated. As mentioned above, P4 pairs found in three *Bacillus* species were likely obtained by lateral gene transfer from the Proteobacteria.

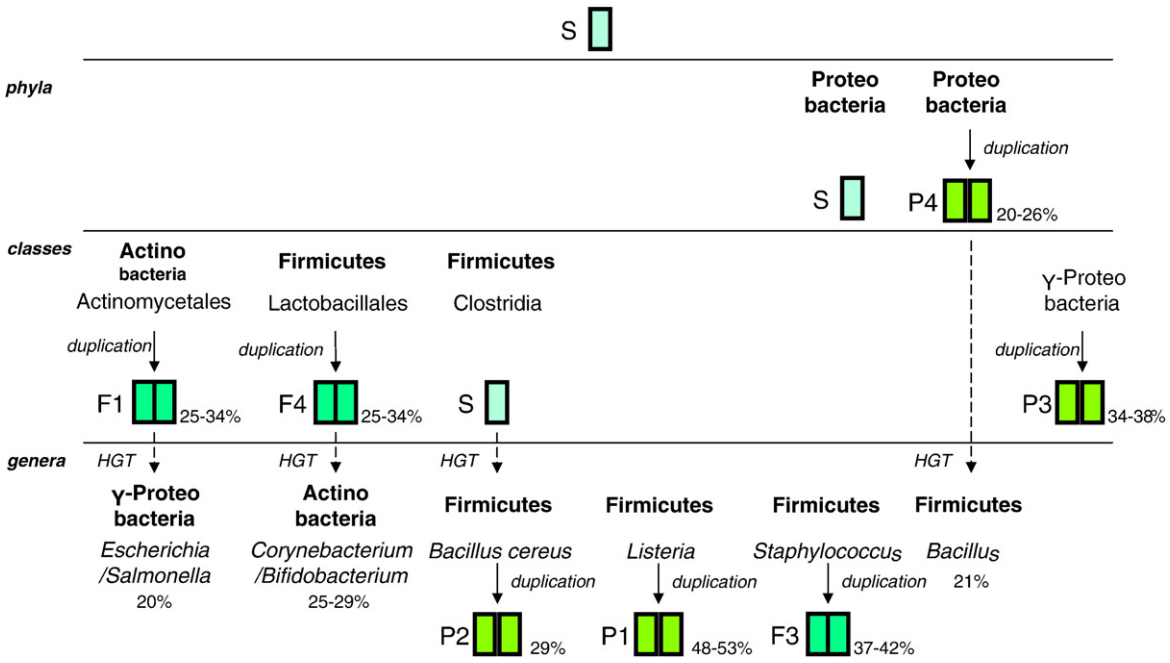

Fig. 7. Reconstruction of the evolution of DUF606 members. Early events are plotted at the top and more recent event towards the bottom. Phyla, classes and genera indicate increasing diversification of the microorganisms. The ancestor single gene is indicated at the top and is assumed to be present in the primordial microorganism. Evolutionary events duplication and HGT (horizontal gene transfer) are indicated. Percentage sequence identities between corresponding proteins in pairs or domains in fusion proteins are given. Subgroups P1–P5 and F1–F4 are defined in Tables 1 and 2. Singles are indicated in light blue, fusions in dark blue, and pairs in green.

Actinobacteria contain almost exclusively fusion proteins, distributed over subgroups F1 and F4. The duplication event resulting in fusions F1 appears to have taken place in a primordial Actinomycetalis. Surprisingly, three F1 fusions also show up in a well defined and limited niche in the γ-Proteobacteria (*Escherichia* and *Salmonella*) where they were probably transferred to laterally. Subgroup F4 contains mostly proteins from Lactobacillales in phylum Firmicutes (see above), but also 3 fusion proteins from the phylum Actinobacteria. The 3 fusions may be the result of a horizontal gene transfer between the two phyla.

Selective advantage

The two-domain membrane proteins with parallel and anti-parallel domain orientation for which a crystal structure has been obtained (see Introduction), appear to be found only in a single state, the fusion state. In these cases, there must have been a strong selective advantage of the fusion proteins over the hetero- and homodimers. In contrast, the DUF606 family is very rich in evolutionary states (Fig. 3), which is exceptional and allowed for the analysis presented here. Selective pressure must have resulted in the evolution of DUF606 members as single, pair or fusion in different organisms. In Firmicutes, the different types of DUF606 proteins evolved largely following the division in classes; singles in Clostridia, fusions in Lactobacillales, and pairs in Bacillales (F3 being the exception), suggesting that the selective advantage was specific for the niches in which these lineages evolved. In the absence of functional information of the DUF606 proteins it is impossible to establish what the advantage could have been. In contrast, in Proteobacteria, pairs evolved alongside singles with no apparent lineage preference, suggesting that a new function evolved with the emergence of pairs that was of special benefit to the organisms. Arguing against such a significantly different function for pairs and singles is that only few organisms have both a pair and a single (paralogues). In general, paralogues (other than the pairs of adjacent genes) are very rare in the DUF606 family with no specific preference for a phylogenetic niche. Selective pressure appears to have driven the evolution of almost exclusively fusions in Actinobacteria.

Because of the large number of evolutionary states still present, the DUF606 family turned out to be very suitable to reconstruct the evolution of a two-domain membrane protein with inverted domain topology. Experimental studies of the function of the different states of the DUF606 members are required to reveal the selective pressure(s) that resulted in the evolution of the different states and more in general of the two-domain membrane proteins.

Computational Methods

Members of the DUF606 family were identified by BLAST searching[20] of a local database containing the translated genes identified on the genomes of 40 archaeal and 490 bacterial strains present in the NCBI microbial database on June 1, 2007†. The set of sequenced genomes contained 330 different bacterial species and 39 archaeal species. BLAST searches were performed using low complexity filtering and composition-based statistics, and a maximal Expect value of 0. The procedure of extracting all members of the family from the database was basically as described.[21] Briefly, all identified members were used as query in the searches and false-positives were filtered out by evaluating hits having Expect values between 10^{-3} and 0 based on hydropathy profile analysis and sequence length. The final dataset was grouped into singles, pairs and fusions as described in the text.

Multiple sequence alignments were done using the command line version of CLUSTAL W[22] for the Windows XP platform that was downloaded from‡. Alternatively, MUSCLE[23] was used but, since the conclusions from both algorithms were the same, only the former was reported. Both programs were used with the default settings. Bootstrap values were calculated using the PHYLIP software package§. In all, 100 random datasets were generated from the input multiple sequence alignments using the program SEQBOOT. Only positions with less than 20% gaps were included. PROTDIST (with the Jones–Taylor–Thornton matrix) and NEIGHBOUR were used to calculate the distances and generating neighbor-joining trees. CONSENSE was used to calculate the consensus tree. Trees were viewed using TreeView.[24]

A consensus topology model was constructed by combining secondary structure prediction by TMHMM2.0 with multiple sequence alignment.[25] All sequences in a group were submitted to the TMHMM predictor at‖, and the gaps observed in each sequence in the multiple sequence alignment of the group were introduced in the predicted topology model and posterior probability profiles. Subsequently, the predictions at each position were averaged. No value was assigned (a gap) when more than 25% of the sequences contained a gap at a position. The positive charge density in the loops at the two sides of the membrane was determined by counting the positive charges (R and K) in the loops defined in the consensus topology model extended by five residues at each side and dividing by the total number of residues in these stretches. N and C termini were treated similarly but, obviously, extended at one side only.

† ftp://ftp.ncbi.nih.gov/_genomes/Bacteria/
‡ ftp://ftp.ebi.ac.uk/pub/software/_dos/clustalw/
§ http://evolution.genetics.washington.edu/phylip.html
‖ http://www.cbs.dtu.dk/services/TMHMM/

Acknowledgements

This work was supported by grants from the Netherlands Organization for Scientific Research (NWO) to DJS and JSL.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.03.005

References

1. Fu, D., Libson, A., Miercke, L. J., Weitzman, C., Nollert, P., Krucinski, J. & Stroud, R. M. (2000). Structural determinants of water permeation through aquaporin-1. *Science*, **290**, 481–486.
2. Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J. B. *et al.* (2000). Structural determinants of water permeation through aquaporin-1. *Nature*, **407**, 599–605.
3. Khademi, S., O'Connell, J., III, Remis, J., Robles-Colmenares, Y., Miercke, L. J. W. & Stroud, R. (2004). Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å. *Science*, **305**, 1587–1594.
4. Van den Berg, B., Clemons, W. M., Jr, Collinson, I., Modis, Y., Hartmann, E., Harrison, S. C. & Rapoport, T. A. (2004). X-ray structure of a protein-conducting channel. *Nature*, **427**, 36–44.
5. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. (2005). Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature*, **437**, 215–223.
6. Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. (2002). X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature*, **415**, 287–294.
7. Hunte, C., Screpanti, E., Venturi, M., Rimon, A., Padan, E. & Michel, H. (2005). Structure of a Na^+/H^+ antiporter and insights into mechanism of action and regulation by pH. *Nature*, **435**, 1197–1202.
8. Locher, K. P., Lee, A. T. & Rees, D. C. (2002). The *E. coli* BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science*, **296**, 1091–1098.
9. Sobczak, I. & Lolkema, J. S. (2005). The 2-hydroxycarboxylate transporter (2HCT) family: physiology, structure and mechanism. *Microbiol. Mol. Biol. Rev.* **69**, 665–695.
10. Lolkema, J. S., Sobczak, I. & Slotboom, D.-J. (2005). Secondary transporters of the 2HCT family contain two homologous domains with inverted membrane topology and trans re-entrant loops. *FEBS J.* **272**, 2334–2344.
11. Sääf, A., Baars, L. & von Heijne, G. (2001). The internal repeats in the Na^+/Ca^{2+} exchanger-related *Escherichia coli* protein YrbG have opposite membrane topologies. *J. Biol. Chem.* **276**, 18905–18907.
12. Poolman, B., Geertsma, E. & Slotboom, D.-J. (2007). A missing link in membrane protein evolution. *Science*, **315**, 1229–1231.
13. Rapp, M., Seppälä, S., Granseth, E. & von Heijne, G. (2007). Emulating membrane protein evolution by rational design. *Science*, **315**, 1282–1284.
14. Rapp, M., Granseth, E., Seppälä, S. & von Heijne, G. (2006). Identification and evolution of dual-topology membrane proteins. *Nature Struct. Mol. Biol.* **13**, 112–116.
15. Schuldiner, S. (2007). When biochemistry meets structural biology: the cautionary tale of EmrE. *Trends Biochem. Sci.* **32**, 252–258.
16. Ubarretxena-Belandia, I., Baldwin, J. M., Schuldiner, S. & Tate, C. G. (2003). Three-dimensional structure of the bacterial multidrug transporter EmrE shows it is an asymmetric homodimer. *EMBO J.* **22**, 6175–6181.
17. Kikukawa, T., Miyauchi, S., Araiso, T., Kamo, N. & Nara, T. (2007). Anti-parallel membrane topology of two components of EbrAB, a multidrug transporter. *Biochim. Biophys. Res. Commun.* **358**, 1071–1075.
18. von Heijne, G. & Gavel, Y. (1988). Topogenic signals in integral membrane proteins. *Eur. J. Biochem.* **174**, 671–678.
19. von Heijne, G. (1989). Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, **341**, 456–458.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
21. Lolkema, J. S. & Slotboom, D. J. (2003). Classification of 29 families of secondary transport proteins into a single structural class using hydropathy profile analysis. *J. Mol. Biol.* **327**, 901–909.
22. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
23. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.
24. Page, R. D. M. (1996). TREEVIEW: an application to display phylogenetic trees on personal computers. *Computer Appl. Biosci.* **12**, 357–358.
25. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.